주제별 분산 지식베이스에 의한 개념기반 정보검색시스템의 성능향상에 관한 연구

A Study on the Improvement of Performance of Concept-Based Information Retrieval Model Using a Distributed Subject Knowledge Base

노영희(Young-Hee Noh)*

초 록

개념기반 정보검색기법은 불리언 검색기법의 문제점을 해소했다고 평가받고 있는 단순 매칭함수 기법이나 P-norm 검색기법보다 높은 성능을 보여주고 있다. 그러나 개념확장에 필수적인 의미망 지식베이스를 구축하는데 시간이 너무 오래 걸리는 단점이 있다. 본 연구에서는 이러한 문제를 해결하기 위해 주제범주별로 지식베이스를 분산 구축함으로써 지식베이스 구축에 소요되는 시간을 단축하면서도 검색성능이 떨어지지 않도록 하는 방안을 모색하고자 하였다.

ABSTRACT

The concept based retrieval model has shown a higher performance than those of the simple matching function method or the P-norm retrieval method introduced to compensate the demerits of the Boolean retrieval model. However, it takes too long to create a semantic-net knowledge base, which is essential in concept exploration. In order to solve such demerits, a method was sought out by creating a distributed knowledge base by subjects to reduce construction time without hindering the performance of retrieval.

키워드 : 개념기반 검색모형, 의미망 지식베이스, 분산 지식베이스, 개념확장, 정보검색
concept based retrieval model, semantic-net knowledge base, distributed knowledge base, concept exploration, information retrieval.

* 이화여대 국제정보센터 실장(trustme@irs4u.net)
■ 논문 접수일 : 2002년 2월 4일
■ 게재 확정일 : 2002년 3월 8일

1 서 론

통계적 정보검색시스템에 대한 연구가 오랫동안 진행되어 오고 있는 한편, 문헌에 출현하는 용어간의 관계를 기반으로 지식베이스를 의미망 구조로 구축하여 놓은 후 이용자의 요구에 대하여 검색을 수행하는 개념기반 정보검색시스템에 대한 연구도 활발하게 진행되어 왔다(Chen et al. 1993; Chen et al. 1994, 노영희 1999).

개념기반 검색모형은 일반적으로 시스템에 입력되는 문헌 데이터베이스로부터 지식베이스를 자동으로 구축하고, 이 지식베이스를 대상으로 개념확장을 수행한 후 문헌 데이터베이스로부터 관련 정보를 검색하는 모형으로서 개념기반 검색모형의 성능을 좌우하는 주요 요소는 지식베이스와 개념확장 알고리즘이라고 할 수 있다.

의미망 구조의 지식베이스를 기반으로 개념확장을 수행하는 알고리즘을 개념확장 알고리즘이라 하며, 개념확장 알고리즘으로는 bnb 알고리즘(branch-and-bound expansion activation algorithm)과 홉필드 넷 알고리즘(Hopfield net algorithm)이 사용되고 있다.

bnb 알고리즘은 적용되는 지식베이스에 따라 크게 경험적(heuristic) bnb 알고리즘과 순차적 bnb(sequential branch-and-bound) 알고리즘으로 구분할 수 있는데, 전자는 주로 전통적인 시소러스에 적용되고 용어간의 관계 정의에 따라 개념확장을 수행한다. 후자는 의미망 구조의 문헌기반 지식베이스에 적용되어 지식베이스 내 용어간의 의미 거리에 따라 개념확장을 수행한다. 그리고 홉필드 넷 알고리즘은 신경망 구조의 지식베이스에 적용되는 개념확장 알고리즘이다.

최근의 한 연구에서는 개념기반 검색기법이 불리언 검색기법의 단점을 어느 정도 해결한 것으로 평가되고 있는 P-norm 검색기법과의 비교실험에서 개념기반 검색기법의 성능이 우수한 것으로 나타났다(노영희 1999). 개념기반 검색모형은 개념확장 대상이 되는 지식베이스로 일반적으로 의미망 구조의 지식베이스를 이용하는데, 이러한 지식베이스를 의미망 구조로 구축하는데 시간이 너무 많이 걸리는 단점이 있다. 따라서 의미망 구조의 지식베이스를 신속하게 구축할 수 있는 방안에 대한 연구가 진행되었으며, 최근에는 UML(Unified Modelling Language) 표기법을 이용하여 지식베이스 및 원문 데이터베이스를 처리함으로써 보다 효과적으로 지식베이스를 구축하고자 하는 연구도 있었다(박병수 2000).

지식베이스를 의미망으로 구축할 때 의미망 구축 시간은 색인어 벡터의 길이가 길수록 오래 걸린다. 따라서 지식베이스를 하나로 구축하지 않고 문헌 집단을 주제분야별로 분류한 후 범주별로 지식베이스를 분산 구축함으로써 의미망 지식베이스 구축 시간을 단축할 수 있을 것이다.

본 연구에서는 이와 같이 전체 문헌

데이터베이스에 대해 하나의 지식베이스를 구축하지 않고 주제 분야별로 분류하여 지식베이스를 분산 구축함으로써 지식베이스 구축 시간을 단축하고자 하였다. 또한 초기 탐색어를 기반으로 개념확장을 할 때, 주제별로 분산 구축된 지식베이스를 대상으로 했을 경우와 모든 데이터를 통합하여 시스템 내 전체문헌을 대상으로 지식베이스를 구축하였을 경우 중 어떤 경우가 더 높은 성능을 보여주는지를 평가하였다.

더 나아가 이렇게 분산 구축된 지식베이스를 대상으로 개념확장을 할 때 이용자의 초기 탐색문에 대하여 개념확장을 할 지식베이스를 어떤 방법으로 선택하는 것이 효율적인지 그 방법론을 고안하고자 하였다.

본 연구의 실험을 위해 사용한 개념확장 알고리즘은 비교적 높은 성능을 보여주고 있는 것으로 평가되고 있는 순차적 bnb 알고리즘이다.

2 이론적 배경

2.1 의미망 지식베이스

개념기반 정보검색에 사용되는 기본적인 지식베이스는 문헌기반 지식베이스로서, 실험 문헌 집단의 각 문헌에 출현한 용어를 자동으로 추출하고 추출된 용어들의 가중치를 산출한다. 또한 용어의 문헌 내 동시출현빈도를 기반으로 용어간의 유사도를 분석하여 의미망 구조의 지식베이스를 최종적으로 구축하게 되는데, 그 과정 및 사용되는 공식을 구체적으로 살펴보면 다음과 같다.

특정 문헌에 출현한 용어의 가중치를 산출하기 위한 공식은 다양하지만 본 연구에서는 단어빈도와 역문헌 빈도를 각각 최대 값으로 나누어 표준화시킨 공식을 사용하였다(Salton, Fox, and Wu 1983).

$$w_{ik} = \frac{tf_{ik}}{\max(tf_{ih})} \times \frac{idf_k}{\max(idf_h)}$$

〈공식 1〉

tf_{ik} = 용어 k가 특정 문헌 i에서 출현한 빈도
$max(tf_{ih})$ = 특정 문헌에서 가장 높은 출현빈도를 갖는 단어의 빈도
idf_k = 용어 k의 역문헌 빈도
$max(idf_h)$ = 전체 문헌 데이터베이스에서 가장 높은 출현빈도를 갖는 단어의 역문헌 빈도

한편, 가중치가 부여된 두 용어간의 의미관계를 생성하기 위해서는 용어간의 유사도가 산출되어야 한다. 개념기반 검색을 위한 의미망 구조의 지식베이스를 구축하는데 사용되는 유사계수는 다양하며 본 연구에서는 코싸인 유사계수를 사용하여 용어간의 유사도를 산출하였다.

$$W(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} \times d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2 \times \sum_{i=1}^{n} d_{ik}^2}}$$

〈공식 2〉

위 공식에서 $W(T_j, T_k)$는 용어 T_j와

용어 T_k간의 유사도 가중치를 나타내고, d_{ij}는 문헌 i에 출현한 용어 T_j의 가중치이며($0 \le d_{ij} \le 1$), d_{ik}는 문헌 i에 출현한 용어 T_k의 가중치이다($0 \le d_{ik} \le 1$).

위와 같이 용어의 추출 및 용어간의 유사도 산출과정을 거친 용어를 기반으로 의미망 구조의 지식베이스를 구축할 수 있다.

2.2 순차적 bnb 알고리즘

의미망 구조의 지식베이스에 적용되는 bnb 확장 활성화 탐색은 개념확장이 진행되는 동안 최단 경로를 찾기 위한 방법이며, 이용자가 제공한 용어에서 개념확장이 시작된다(Chen and Dhar 1991). 이용자가 제공한 초기 탐색어에는 1의 가중치가 부여되며, 다음으로 이 용어들과 직접적으로 관련이 있는 이웃한 용어들을 탐색한다. 확장된 용어의 가중치는 이용자가 입력한 용어와의 링크 가중치를 기반으로 산출된다. 순차적 bnb 알고리즘의 개념확장 과정을 단계적으로 기술해 보면 다음과 같다.

(1) 이용자가 탐색문을 입력한다. 이용자가 입력한 초기 탐색어 집합이 {S_1, S_2, ... S_m}일 때, 의미망에 나타난 용어들 중 초기 탐색어와 일치하는 용어는 1의 가중치를 갖는다.

$$\mu_i(0) = x_i,\ 0 \le i \le n\text{-}1$$

$\mu_i(t)$는 t번 반복한 후의 노드 i의 가중치이다. 초기 탐색어에 할당된 노드의 가중치는 1이다.

(2) 순차적 bnb 알고리즘은 내림차순으로 우선순위 대기행렬인 $Q_{priority}$를 생성한다. 최초의 우선순위 대기행렬은 아래와 같다.

$$Q_{priority} = \{ S_1, S_2, \ldots S_m \}$$

또한, 출력 대기행렬인 Q_{output}을 생성해야 하는데, 이는 확장이 반복되는 동안 활성화 노드를 저장하기 위해서이다.

$$Q_{output} = \{\ \}$$

(3) 반복이 계속되는 동안, bnb 알고리즘은 $Q_{priority}$에서 가장 높은 가중치의 노드들을 제거하고 그들의 이웃 노드들을 활성화시키며, 다음 공식에 의해 이웃 노드들의 가중치를 산출한다.

$$\mu_j(t+1) = \mu_j(t) \times t_{ij} \qquad \langle 공식\ 3 \rangle$$

위 식에서 $\mu_j(t+1)$는 bnb 알고리즘에 의해 확장될 새로운 노드의 가중치이고, $\mu_j(t)$는 확장 전 노드의 가중치이며 t_{ij}는 확장 전 노드와 확장될 새로운 노드의 유사도 가중치 즉, 링크 가중치이다. 새롭게 활성화될 노드의 가중치는 활성화될 노드와 활성화되기 전 노드간의 링크 가중치에 의존한다.

(4) 활성화되었던 노드는 출력 대기행렬, Q_{output}에 저장된다. 계산이 끝난 후 모든 활성 노드들은 가중치순으로 정렬되어 $Q_{priority}$에 저장된다.

(5) 두 개의 다른 노드로부터 도달되는 노드의 가중치는 두 노드간의 유사도 가중치를 합하여 산출할 수 있다. 이와 같이 의미망에서 두 개의 다른 시작 노드에 의해 도달되어질 수 있는 하나의 노드에 보다 높은 가중치를 할당하는 기법은 기타 다른 확장 활성화 탐색에 채택되어 왔다(Shoval 1985; Cohen and Kjeldsen 1987; Chen and Dhar 1991).

2.3 개념확장 대상 지식베이스 선정

전체 문헌을 대상으로 했을 때와 달리 개념기반 정보검색에 사용될 지식베이스를 구축할 때 전체 문헌 집단을 대상으로 하거나 문헌 집단을 주제별로 분산 구축할 수 있다. 문헌 집단을 주제별로 분류한 후 의미망 지식베이스를 분산 구축한 경우에는 여러 개의 주제별 지식베이스가 있기 때문에 이용자가 입력한 초기 탐색어를 기반으로 개념확장을 할 지식베이스를 선정해야 한다. 즉 초기 탐색문에 출현한 각각의 탐색어들에 대하여 주제별 지식베이스 중 어느 지식베이스를 대상으로 개념확장을 함으로써 최종 탐색문을 완성할 것인지를 결정해야 하는 것이다. 본 연구에서는 크게 네 가지 방법에 의해서 개념확장 대상 지식베이스를 선정하고자 하였다.

2.3.1 검색문헌 수의 활용

이 방법은 초기 탐색문으로 1차 문헌검색을 한 검색결과를 지식베이스 선정에 활용하는 방법이다. 즉, 이용자가 제공한 초기 탐색문으로 전체 문헌 데이터베이스를 대상으로 1차 문헌 검색을 한 후 상위 특정 순위까지의 문헌을 분석한다. 다음으로 검색된 문헌들이 가장 많이 출현한 지식베이스를 선정하여 그 선정된 지식베이스를 대상으로 개념확장을 한 후 최종 탐색문을 완성하는 방법이다. 이 때 1차 검색을 하여 검색된 결과의 순위화에 사용한 공식은 코싸인 유사계수 공식이다. 이를 간단히 공식으로 표현하면 공식 4와 같다.

$$KB_x = \sum_{i=1}^{n} RES_k \qquad \langle 공식\ 4 \rangle$$

위 공식에서 KB_x는 이용자의 탐색문을 기반으로 산출된 각 지식베이스의 적합성 가중치를 나타내고, x는 주제별로 분산 구축된 지식베이스의 총 수이다. n은 초기 탐색문을 기반으로 검색된 문헌의 수이고, RES는 검색된 문헌이며, k는 검색된 문헌 중 각 지식베이스에 속한 문헌이다.

위와 같은 방법을 사용할 경우 상위 순위에 있는 문헌들이 가장 많이 출현한 지식베이스를 선정할 때 동순위 문제가 발

생할 수 있다. 동순위 문제를 해결하는 방법으로서 순위화된 결과 중 동순위에 해당하는 지식베이스에 출현한 문헌들의 유사도 가중치를 합하여 더 높은 가중치를 갖는 지식베이스를 선택하는 방법이 있다.

2.3.2 검색결과 중 유사도 가중치의 합 활용

이 방법은 검색문헌 수를 활용하는 방법과 유사하지만 검색문헌 수가 아니라 검색된 문헌들의 유사도 가중치를 활용하여 지식베이스를 선정하는 점이 다르다. 즉, 검색문헌 수를 활용하는 방법과 유사하게 초기 탐색문에 대하여 전체 문헌 집단을 대상으로 1차 문헌 검색을 한 후 상위 순위에 있는 문헌들의 유사도 가중치의 합이 가장 높은 지식베이스를 선정하는 방법이다. 이와 같이 유사도 가중치를 이용할 경우 검색 문헌 수를 활용할 때 발생했던 동순위 문제가 발생하지 않는다. 이를 간단하게 공식으로 표현하면 공식 5와 같다.

$$KB_x = \sum_{i=1}^{n} RES_{kw} \qquad \langle 공식\ 5 \rangle$$

위 공식에서 kw는 검색된 문헌 중 각 지식베이스에 속한 문헌들의 유사도 가중치이며, 유사도 가중치를 산출하기 위해 사용된 공식은 코싸인 유사계수이다.

2.3.3 탐색어의 출현빈도 활용

이 방법은 초기 탐색어의 전체 데이터베이스 내 문헌빈도(Document Frequency)를 활용하는 방법이다. 즉 초기 탐색문을 이루는 각 탐색어의 문헌빈도(DF)를 지식베이스별로 산출한 후 가장 높은 문헌빈도를 갖는 지식베이스를 개념확장 대상 지식베이스로 선정하는 방법이다. 이를 공식으로 표현하면 공식 6과 같다.

$$KB_x = \sum_{i=0}^{l} DF_{iq} \qquad \langle 공식\ 6 \rangle$$

위 공식에서 DF_{iq}는 탐색문을 이루는 각 탐색어의 문헌빈도이고 l은 탐색문을 이루는 탐색어의 총 수이다.

2.3.4 탐색어의 출현빈도 및 확장용어 가중치의 활용

이 방법은 초기 탐색어의 문헌빈도 뿐만 아니라 초기 탐색어에 의해 확장된 용어들의 확장용어 가중치를 모두 활용하는 방법이다. 즉, 초기 탐색문을 이루는 각 탐색어의 문헌빈도(DF)를 산출하는 한편, 초기 탐색문에 대하여 각 지식베이스를 대상으로 개념확장 조건을 만족할 때까지 개념확장을 한 후 확장된 용어들과의 유사도 가중치를 합한다. 그런 다음 가장 높은 유사도를 갖는 지식베이스를 선택하여 문헌검색을 하는 방법이다. 이를 간단하게 공식으로 표현하면 공식 7과 같다.

$$KB_x = \sum_{i=1}^{l} DF_{iq} \times \sum_{i=0}^{m} ET_{xw} \qquad \langle 공식\ 7 \rangle$$

위 공식에서 ET_{xw}는 개념확장에 의해 확장된 용어의 가중치(the weight of the extended term)이며, m은 확장된 용어의 수로서 보통 이용자가 탐색문을 제공할 때 확장될 최대 용어 수를 제공한다.

3 실험 설계

3.1 문서 수집 및 분류

본 연구에서는 실험데이터로서 웹 로봇 에이전트를 이용하여 수집된 인터넷 문서를 활용하고 있다. 수집된 문서의 주제분야는 경제학분야이다. 수집된 문서들은 지식베이스의 주제별 분산 구축의 효율성을 알아보기 위해 몇 개의 주제분야로 분류되는데, 문서의 주제별 분류를 위해 사용한 분류체계는 DDC 분류체계이며 330에서 339까지 크게 10구분하였다.

실험의 공정한 평가를 위해 주제전문가 2명이 총 10개의 탐색문을 실험 문헌 집단에 출현한 용어 중에서 추출하였으며, 이 탐색문들은 모두 3~5개의 탐색어로 구성되어 있다.

3.2 지식베이스 구축

본 연구에서는 전체 문헌 데이터베이스를 대상으로 지식베이스를 구축한 경우와 주제별로 분산시켜 지식베이스를 구축한 경우 중 어느 것이 보다 더 높은 성능을 보여주는지를 평가하기 위해 두 가지 방법으로 의미망 지식베이스를 구축하였다. 이를 위해 DDC 분류체계 330에서 339까지 분류번호에 따라 10개의 지식베이스를 주제별로 분산 구축하였고 각 분류번호에 분류된 문서를 모두 통합하여 문서 전체에 대한 지식베이스를 크게 하나로 구축하였다.

3.3 실험 조건 및 평가척도

실험 문헌 데이터베이스의 총 건수는 4,421건으로서 이 문서들은 웹로봇에 의해 수집되어 중분류(330~339) 주제범주에 자동으로 분류된 것이다.

개념확장 조건으로 최대 확장용어 수와 확장될 용어의 최저 가중치는 선행 연구에서 비교적 높은 성능을 보여 주었던 조건만을 취하였다(노영희, 정영미 2000). 즉, 최대 확장용어 수는 8개로 하였고 확장될 용어의 최저 가중치는 0.4로 하였으며, 검색된 결과의 평가는 상위 10, 20, 30, 50 순위 까지의 문서를 대상으로 하고 있다.

성능을 평가하기 위한 척도로서 가장 일반적으로 이용되는 재현율(p)과 정확률(r)을 사용하였으며, 반비례 관계에 있는 재현율과 정확률을 모두 반영하기 위해 사용되는 척도로 van Rijsbergen (1979)에 의해 처음으로 도입된 F척도(F-measure)를 사용하되(Moulinier, Raskinis, and Ganascia 1996; Lewis et al. 1996; Cohen and Singer 1996) 정확률과 재현율을 하나의 값으로 나타나기 위하여 β값을 1로 하는 F_1척도를 부분적으로 사

용하였다. F_1척도를 부분적으로 사용한 이유는 재현율은 높은데 정확률이 낮거나 그 반대인 경우, 정확한 성능평가를 하기 위함이었다.

$$F_\beta(r,p)=\frac{(\beta^2+1)pr}{\beta^2 p+r}$$

$$F_1(r,p)=\frac{2pr}{p+r}$$

4 실험결과 분석

4.1 의미망 지식베이스 구축 시간

지식베이스를 의미망으로 구축할 때 의미망 구축 시간은 색인어 벡터의 길이가 길수록 오래 걸린다. 따라서 본 연구에서는 의미망 지식베이스를 하나로 구축하지 않고 문헌을 주제분야별로 분산 구축하여 색인어 벡터의 길이를 줄임으로써 의미망 구축 시간을 단축하고자 하였다. 지식베이스 구축을 위해 사용한 시스템은 Compaq의 Alpha 4100 Digital Unix로서 1CPU, Memory 512M이다. 실험이 진행되는 시스템에 지식베이스 구축 프로세스 이외의 다른 프로세스가 수행되지 않도록 통제하였으며 이는 시스템의 다른 프로세스가 수행됨으로써 실험 결과에 영향을 미칠 수 있기 때문이었다.

〈표 1〉과 〈표 2〉는 주제범주별로 문헌을 분류한 후 각 범주별로 지식베이스를 분산 구축한 경우와 전체 문헌 데이터베이스를 대상으로 지식베이스를 구축한 경우의 구축 시간을 비교한 것이다. 〈표 1〉에 나타난 지식베이스 구축 시간은 10개의 범주에 대한 지식베이스를 동시에 돌려서 병렬적

〈표 1〉 주제범주별로 분류된 문서 수

분류 번호	분류된 문서 수	의미망 구축 시간
330	105	10분
331	319	23분
332	1335	92분
333	345	32분
334	13	6분
335	33	12분
336	1006	79분
337	40	11분
338	1081	70분
339	144	15분
총	4,421	350분/35.00분(평균)

〈표 2〉 전체 문서 집단의 문서 수

	전체 문서 수	의미망 구축 시간
전체문서_KB	4,421	302분

으로 구축할 때 걸린 시간이다. 표에서 보듯이 주제범주별로 문헌을 10개의 범주로 분류하고 10개 범주의 데이터베이스에 대해 각각의 지식베이스를 구축한 경우 지식베이스 총 구축 시간은 350분이지만 구축작업이 병렬적으로 수행되었기 때문에 의미가 없고 오히려 평균 구축 시간 또는 10개의 범주 중 구축 시간이 가장 긴 지식베이스 구축 시간과 비교하는 것이 보다 의미 있는 것으로 판단된다.

먼저, 각 지식베이스의 평균 구축 시간과 비교한 경우에 구축 시간은 35.00분으로서 전체 문헌 대상 지식베이스 구축 시간, 302분보다 약 267분, 8.6배 정도 단축되었다. 다음으로 10개의 주제 범주 중 지식베이스 구축이 가장 늦게 완료된 332 분류번호의 지식베이스의 구축 시간은 92분으로서 전체 문헌 대상 지식베이스 구축 시간보다 약 210분, 3.3배 정도 단축되었다.

〈표 1〉에서 분류된 문서 수에 대응한 의미망 구축 시간을 볼 때, 문서 수가 많아질수록 의미망 구축 시간이 길어지는 것으로 보아 문서 4,421건에 대한 302분의 의미망 구축 시간은 당연한 결과인 것으로 보이며, 따라서 지식베이스 구축 시간을 단축하기 위해 문헌을 어떤 방식으로든 분류하여 지식베이스를 분산 구축하는 것은 의미있는 일로 분석된다.

4.2 1차 문헌검색을 통한 지식베이스 선정 결과

지식베이스의 분산 구축을 기반으로 한 문헌검색의 성능을 평가하기 전에 이용자의 초기 탐색어를 기반으로 개념확장 대상 지식베이스를 선정할 수 있는 방법들에 대하여 평가할 필요가 있다. 먼저, 분산 구축된 지식베이스 중 적절한 개념확장 대상 지식베이스를 선정하는 방법 중 초기 탐색문으로 전체 문헌 집단을 대상으로 1차 문헌 검색을 한 후 상위 순위에 있는 문헌들이 가장 많이 출현한 지식베이스를 선정하는 방법(공식 4 참조)과 초기 탐색문으로 전체 문헌 집단을 대상으로 1차 검색을 한 후 상위 순위에 속한 문헌들의 유사도 가중치의 합이 가장 높은 지식베이스를 선정하는 방법(공식 5 참조)을 비교하였다.

〈표 3〉과 〈표 5〉는 전체 문헌 집단을 대상으로 1차 검색을 한 후 검색 결과 중 상위 10순위 및 20순위에 각각 속한 문헌들의 통계를 보여 주고 있고 〈표 4〉와 〈표 6〉은 동일한 방법으로 검색된 결과 중 상위 10순위 및 20순위에 각각 속한 문헌들의 유사도 가중치의 합을 보여 주고 있다.

표에서 보는 바와 같이 개념확장 대상 지식베이스를 선택할 때 검색된 문헌 중 상위에서 발견된 문헌들의 수를 기반으로 하든지 아니면 상위 문헌들의 유사도 가중치의 합으로 하든지 선택된 주제범주가 거의 일치한다는 것을 알 수 있다. 다만 검색된 문헌들의 수를 기반으로 할 경우

〈표 3〉 검색된 문헌 중 10위까지의 문헌 수 통계

탐색문 / 주제범주	1	2	3	4	5	6	7	8	9	10
330	0	0	1	0	1	9	0	0	0	0
331	2	2	0	1	1	0	7	3	0	3
332	4	6	1	1	1	0	0	5	3	0
333	1	0	0	2	0	0	0	0	1	2
334	0	0	0	0	0	0	0	0	0	0
335	0	0	0	0	0	0	0	0	0	0
336	1	0	2	1	4	1	3	1	5	3
337	0	0	0	2	0	0	0	0	0	0
338	1	2	5	3	3	0	0	1	1	2
339	1	0	1	0	0	0	0	0	0	0

〈표 4〉 검색된 문헌 중 10위까지의 문헌들의 유사도 가중치의 합

탐색문 / 주제범주	1	2	3	4	5	6	7	8	9	10
330	0.0000	0.0000	0.8603	0.0000	0.9869	7.7760	0.0000	0.0000	0.0000	0.0000
331	1.7931	1.6580	0.0000	0.9737	0.9778	0.0000	5.1579	2.4627	0.0000	2.7222
332	3.4712	5.2090	0.8625	0.9901	0.9725	0.0000	0.0000	4.3397	2.8170	0.0000
333	0.8125	0.0000	0.0000	1.9503	0.0000	0.0000	0.0000	0.0000	0.8603	1.7381
334	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
335	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
336	0.8549	0.0000	1.7253	0.9748	3.8854	0.8895	2.6484	0.8651	4.6950	2.8440
337	0.0000	0.0000	0.0000	1.9255	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
338	0.8433	1.6526	4.3092	2.9222	2.9371	0.0000	0.0000	0.8028	0.9652	1.8140
339	0.8457	0.0000	0.8658	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

〈표 5〉 검색된 문헌 중 20위까지의 문헌 수 통계

탐색문 / 주제범주	1	2	3	4	5	6	7	8	9	10
330	0	0	1	0	1	19	0	0	1	0
331	3	2	0	2	3	0	7	8	0	3
332	6	11	7	5	3	0	10	7	3	3
333	4	0	1	2	0	0	0	0	3	5
334	0	0	0	0	0	0	0	0	0	0
335	0	0	0	0	0	0	0	0	0	0
336	4	4	4	2	8	1	3	3	8	7
337	0	1	0	2	0	0	0	0	0	0
338	2	2	6	5	4	0	0	2	4	2
339	1	0	1	2	1	0	0	0	1	0

〈표 6〉 검색된 문헌 중 20위까지의 문헌들의 유사도 가중치의 합

탐색문 / 주제범주	1	2	3	4	5	6	7	8	9	10
330	0.0000	0.0000	0.8603	0.0000	0.9869	16.4160	0.0000	0.0000	0.8164	0.0000
331	2.5588	1.6580	0.0000	1.9034	2.8173	0.0000	5.1579	6.2447	0.0000	2.7222
332	5.0559	9.2884	6.0090	4.6731	2.8631	0.0000	7.7100	5.8555	2.8170	2.5734
333	3.1096	0.0000	0.8575	1.9503	0.0000	0.0000	0.0000	0.0000	2.7666	4.3115
334	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
335	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
336	3.2018	3.2642	3.4395	1.8851	7.6060	0.8895	2.6484	2.4098	7.4072	6.2752
337	0.0000	0.8164	0.0000	1.9255	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
338	1.6090	1.6526	5.1682	4.7562	3.8908	0.0000	0.0000	1.5481	3.7822	1.8140
339	0.8457	0.0000	0.8658	1.9036	0.9178	0.0000	0.0000	0.0000	0.8817	0.0000

동순위 문제가 발생한다. 10위까지의 문헌들을 분석한 〈표 3〉의 탐색문 10을 보면 331과 336의 2개의 범주에서 검색된 문헌 수가 3으로 동순위가 발생한다. 이러한 경우에는 동순위 가운데 어느 주제범주를 대상으로 개념확장을 해야할 것인지 선택하여야 한다.

그러나 이러한 동순위 문제는 검색된 문헌들의 범주별 유사도 가중치의 합을 근거로 해결의 실마리를 찾을 수 있다. 즉, 331범주와 336범주에서 검색된 문헌은 모두 3개이지만 검색된 문헌들의 유사도 가중치의 합은 각각 2.7222와 2.8440으로서 336범주가 선택된다.

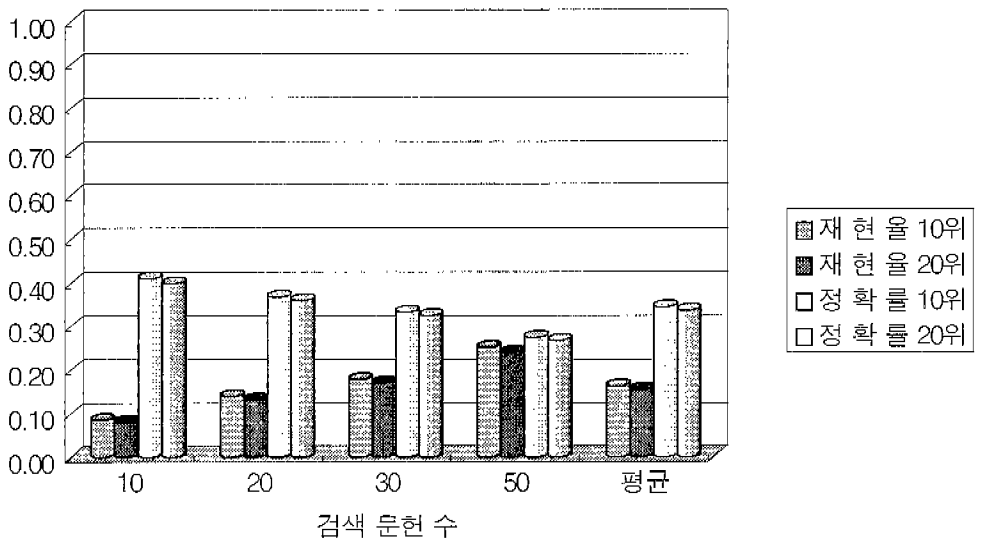
이와 같이 검색된 문헌의 수는 같아도 유사도 가중치의 합이 다른 것은 검색 순위가 다르기 때문이며 결과적으로 높은 검색 순위에 있는 문헌들이 속한 범주가 선택될 가능성이 높아진다.

지금까지의 실험결과를 통해 1차 문헌 검색에 의해 검색된 문헌들을 기반으로 개념확장 대상 지식베이스를 선택할 때 검색된 문헌의 수보다는 유사도 가중치의 합을 기준으로 선택하는 것이 바람직하다는 알 수 있었다. 이 때 상위 몇 순위까지의 검색문헌 수를 분석해서 개념확장 대상 지식베이스를 선택하는 것이 더 높은 성능을 보여 주는지를 비교할 필요가 있다.

〈표 7〉은 유사도 가중치의 합을 기준으로 지식베이스를 선택할 때 1차 검색결과의 상위 10순위와 20순위까지를 분석하였을 때의 재현율 및 정확률의 성능을 평가한 것이고 그림 1은 이를 그림으로 표현한 것이다. 표에서 볼 수 있듯이 1차적으로 검색된 문헌 중 상위 10순위에 속한 문헌만을 분석하였을 경우가 20순위까지를 분석하였을 경우보다 재현율과 정확률에 있어서 각각 0.01%, 0.008%정도 약간 높게 나타났다. 결과에 따르면 분산 구축된 지식베이스들 가운데 특정 질문에 대한 적절한 지식베이스를 선택하기 위해 1차

〈표 7〉 검색 순위별 재현율 및 정확률 평가

1차 검색결과 평가대상 / 검색 문헌 수	재 현 율		정 확 률	
	10위	20위	10위	20위
10	0.0857	0.0824	0.4100	0.4000
20	0.1401	0.1320	0.3700	0.3600
30	0.1778	0.1697	0.3333	0.3267
50	0.2531	0.2403	0.2740	0.2680
평 균	0.1642	0.1561	0.3468	0.3387



〈그림 1〉 검색 순위별 재현율 및 정확률 평가

검색결과를 분석할 때, 20순위까지의 문헌을 모두 고려하든 아니면, 10순위까지의 문헌들만을 고려하든 큰 차이는 없는 것으로 보인다.

4.3 분산 지식베이스 선정 방법에 따른 성능차이

본 절에서는 분산 구축된 지식베이스 중 이용자의 초기 탐색문에 대한 개념확장 대상 지식베이스를 선정하는 방법에 따른 성능차이를 분석하였다. 이용자의

초기 탐색문을 기반으로 한 개념확장 대상 지식베이스를 선정하는 방법으로 네 가지 방법을 고안하고 그 성능 비교를 하였다. 첫 번째 방법은 이용자의 초기 탐색문을 기반으로 전체 문헌 데이터베이스를 대상으로 1차 검색을 한 결과를 이용하여 개념확장 대상 지식베이스를 선정한 후 선정된 지식베이스만을 대상으로 최종 검색을 하는 방법이고, 두 번째 방법은 1차 검색 결과를 활용하여 지식베이스를 선정하여 개념확장을 하되 전체 문헌 데이터베이스를 대상으로 최종 검색을 하는 방법이다. 첫 번째 방법과 두 번째 방법을 사용함에 있어 앞 절에서의 실험결과를 활용하고 있다. 즉, 앞 절에서는 1차 검색 결과 중 검색된 문헌 수를 활용하는 방법과 검색된 문헌 중 유사도 가중치의 합을 활용하는 방법을 비교하였을 때 유사도 가중치를 활용하는 것이 바람직한 것으로 분석되었고, 또한 검색결과 중 상위 몇 건을 활용하는 것이 더 높은 검색효율을 나타낼 것인지에 대한 실험에서는 상 하위간에 별 다른 차이가 없는 것으로 나타나고 있으나, 상위 10위까지의 문헌이 오히려 상위 20위까지의 분석결과를 적용하는 것보다 더 높은 검색효율을 나타냄을 보여주고 있다. 따라서 위의 두가지 방법은 검색된 결과 중 상위 10위까지 문헌의 유사도 가중치의 합을 이용하고 있다.

실험의 세 번째 방법은 각 지식베이스에 출현한 탐색어의 출현빈도를 활용하는 방법(공식 6 참조)으로서 이용자가 입력한 초기 탐색문의 탐색어의 문헌빈도가 가장 높은 지식베이스를 선정하여 개념확장을 하는 방법이고 네 번째 방법은 초기 탐색어를 기반으로 확장된 용어의 가중치, 즉 확장용어 가중치를 탐색어의 문헌빈도에 곱하여 그 값이 가장 높은 지식베이스를 선정하는 방법(공식 7 참조)이다. 〈표 8〉과 〈표 9〉는 각각 분산 지식베이스 선정 방법에 따른 재현율과 정확률 차이를 나타내고 있고 〈그림 2〉와 〈그림 3〉은 이를 그림으로 표현한 것이다.

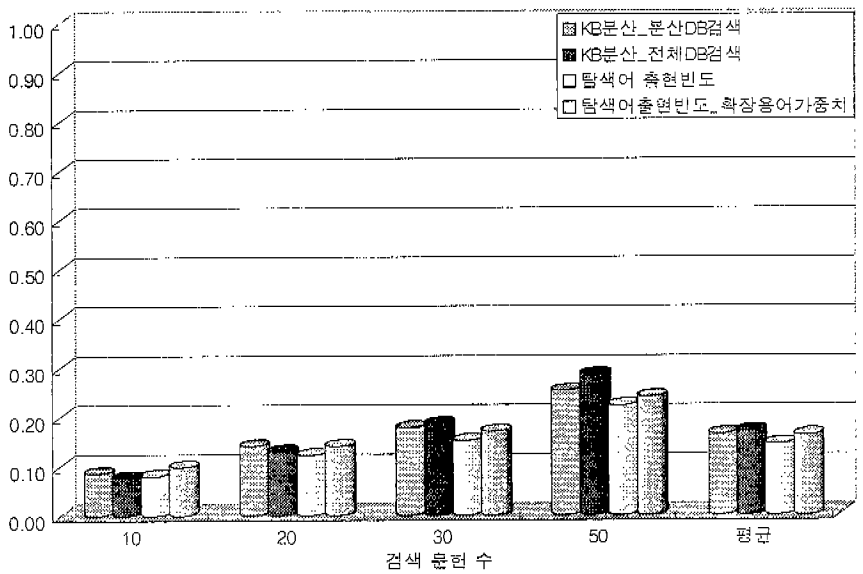
실험 결과에 의하면 재현율에 있어서는 1차 검색결과를 기반으로 하되 선택된 지

〈표 8〉 분산 지식베이스 선정 방법에 따른 재현율

KB 선정방법 / 검색문헌수	재 현 율			
	1차 검색__분산DB검색	1차 검색__전체DB검색	탐색어 출현빈도	탐색어출현빈도__확장용어가중치
10	0.0857	0.0777	0.0802	0.0985
20	0.1401	0.1302	0.1218	0.1401
30	0.1778	0.1856	0.1517	0.1700
50	0.2531	0.2858	0.2222	0.2406
평균	0.1642	0.1698	0.1440	0.1623

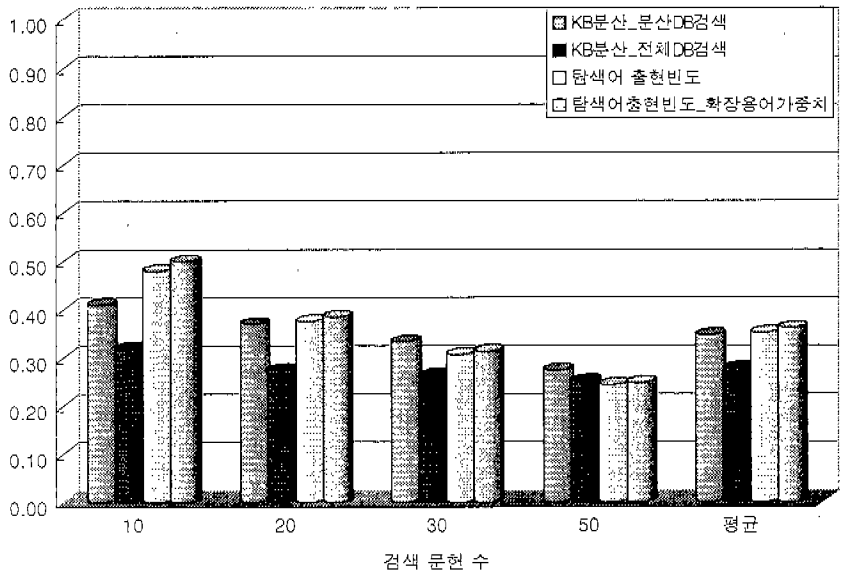
〈표 9〉 분산 지식베이스 선정 방법에 따른 정확률

KB 선정방법 / 검색문헌 수	정 확 률			
	1차 검색__ 분산DB검색	1차 검색__ 전체DB검색	탐색어 출현빈도	탐색어출현빈도__ 확장용어가중치
10	0.4100	0.3200	0.4800	0.5000
20	0.3700	0.2750	0.3750	0.3850
30	0.3333	0.2633	0.3067	0.3133
50	0.2740	0.2520	0.2440	0.2480
평균	0.3468	0.2776	0.3514	0.3616



〈그림 2〉 분산 지식베이스 선정 방법에 따른 재현율

식베이스를 대상으로 개념확장을 한 후 전체 문헌 데이터베이스를 검색하였을 때의 성능이 가장 높은 것으로 나타났고, 탐색어의 출현빈도만을 기반으로 할 때의 성능이 가장 낮은 것으로 나타났다. 반면에 정확률은 탐색어의 출현빈도에 확장용어 가중치를 반영한 경우가 가장 높은 성능을 보여 주었고, 1차 검색결과를 기반으로 하되 선택된 지식베이스를 대상으로 개념확장을 한 후 전체 문헌 데이터베이스를 검색하였을 때의 성능이 가장 낮은 것으로 나타났다. 후자의 경우 전체문헌을 대상으로 검색하므로 재현율은 높지만 검색된 문헌 중 적합 문헌이 상위에서 발



〈그림 3〉 분산 지식베이스 선정 방법에 따른 정확률

견되지 못함으로써 정확률이 낮아지는 것으로 분석된다.

〈표 10〉은 분산 지식베이스 선정방법에 따른 F척도 값을 비교한 것으로 재현율과 정확률에 동일한 비중을 두어 하나의 값으로 표현하여 그 성능을 비교하였다. 그 결과를 보면 탐색어의 출현빈도에 확장용어 가중치를 반영한 경우의 F척도 값이 평균 0.2240%로 가장 높은 성능을 보여 주었고 탐색어의 출현빈도만을 고려한 경우의 F척도 값은 평균 0.2043%로 가장 낮았다.

1차 검색 후 검색대상 문헌 집단을 분산 데이터베이스로 하느냐 전체 문헌 데이터

〈표 10〉 분산 지식베이스 선정 방법에 따른 F척도 비교

KB 선정방법 / 검색문헌수	F 척도			
	1차 검색__분산DB검색	1차 검색__전체DB검색	탐색어 출현빈도	탐색어출현빈도__확장용어가중치
10	0.1418	0.1250	0.1374	0.1646
20	0.2032	0.1767	0.1839	0.2054
30	0.2319	0.2177	0.2030	0.2204
50	0.2631	0.2678	0.2326	0.2442
평균	0.2229	0.2107	0.2043	0.2240

〈표 11〉 지식베이스 구축방법에 따른 재현율 및 정확률 평가

성능 / 검색 문헌 수	재현율			정확률		
	개념확장전	개념확장후		개념확장전	개념확장후	
		KB분산__Qdf__ETw	전체DB__KB		KB분산__Qdf__ETw	전체DB__KB
10	0.1095	0.0857	0.1333	0.3800	0.4100	0.4700
20	0.1708	0.1401	0.1947	0.3200	0.3700	0.3900
30	0.2042	0.1778	0.2437	0.2700	0.3333	0.3433
50	0.3001	0.2531	0.3350	0.2500	0.2740	0.2780
평균	0.1961	0.1642	0.2267	0.3050	0.3468	0.3703

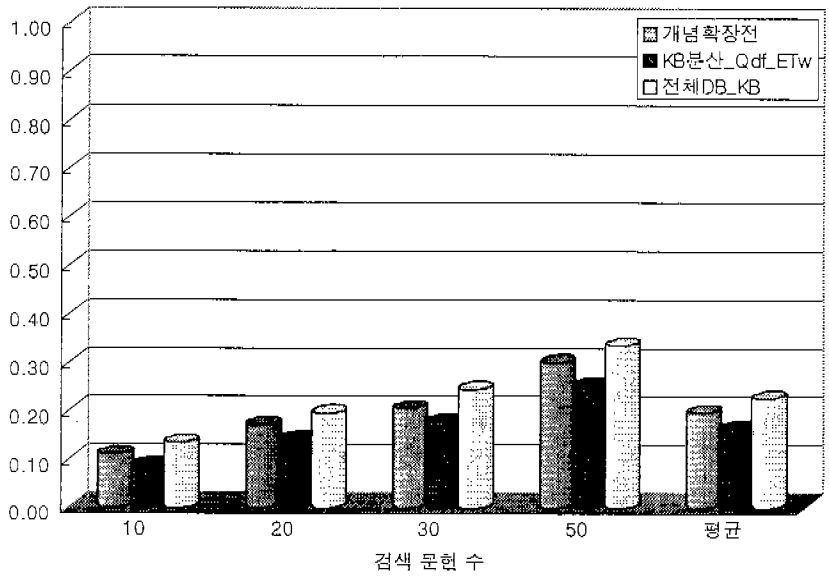
베이스를 대상으로 하느냐에 따라 성능이 어느 정도 차이가 나는지를 F척도를 사용하여 비교하여 보았을 때, 선정된 분산 데이터베이스를 대상으로 하는 것이 전체 문헌 데이터베이스를 하는 경우보다 평균적으로 약 0.0122% 높게 나타났다. 이는 이용자의 초기 탐색문을 기반으로 선정된 분산 데이터베이스 안에 대부분의 적합 문헌이 포함되어 있음을 의미한다.

4.4 지식베이스 구축방법에 따른 성능 차이

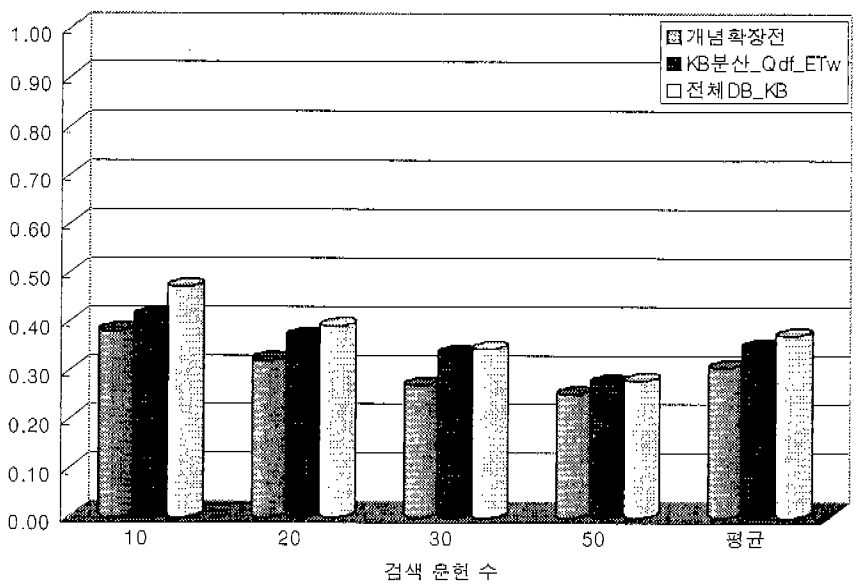
본 절에서는 주제범주별 지식베이스 분산 구축의 효율성을 알아보기 위해 크게 세 가지 경우로 구분하여 성능을 비교하였다. 먼저 이용자의 초기탐색어에 대하여 개념확장을 하지 않고 단순 문헌 검색을 한 경우(개념확장 전)와 분산 구축된 지식베이스 또는 전체문헌기반 지식베이스를 대상으로 개념확장 후 문헌검색을 한 경우를 비교하였다. 개념확장 방법을 사용하는 후자의 방법은 다시 지식베이스를 분산 구축한 경우(KB분산__Qdf__ETw)와 전체문헌을 기반으로 지식베이스를 구축한 경우(전체DB__KB)로 구분하여 실험하였다. 지식베이스를 분산 구축한 경우는 앞 절의 실험에서 성능이 가장 높게 나타난 경우(공식 7 참조)를 선택하여 비교하였다.

〈표 11〉은 지식베이스 구축방법에 따른 재현율과 정확률의 성능 차이를 보여 주고 있고 〈그림 4〉와 〈그림 5〉는 이를 그림으로 표현한 것이다. 표에서 보듯이 평균으로 보았을 때 전체 문헌 데이터베이스를 대상으로 지식베이스를 구축하고 이를 대상으로 개념확장을 한 경우가 재현율과 정확률에 있어서 보다 높은 성능을 보여주고 있음을 알 수 있다.

개념확장을 한 경우와 하지 않은 경우를 비교했을 때에는 개념확장 전보다 개념확장 후가 보다 높은 성능을 보여주고 있었다. 지식베이스를 분산 구축한 경우와



〈그림 4〉 지식베이스 구축방법에 따른 재현율 평가



〈그림 5〉 지식베이스 구축방법에 따른 정확률 평가

〈표 12〉 질문 별 최적의 성능 주제범주

검색문헌 수 \ 탐색문	1	2	3	4	5	6	7	8	9	10
10	332	332	336	336 338	330 332	333 338	336	332 336	332	336
20	339	332	331 336	338	330 332 338	338	336	332 336	332	336
30	339	332	331 336	338	330 332 338	338	336	332 336	336	336
50	338	332	338	338	330 332 336 338	336	336	332 336	336	336
최적의 주제범주	339	332	336	338	330 332	338	336	332 336	332 336	336

단순 문헌 검색을 한 경우를 비교했을 때에 재현율은 단순 문헌 검색을 한 경우가 0.0319% 높지만 정확률은 0.0418% 낮은 것으로 나타났다.

위의 실험결과에서 볼 때 지식베이스를 분산 구축하는 것보다 통합 구축하는 것이 재현율과 정확률에 있어 각각 0.0625%, 0.0235% 높은 것으로 나타났다. 이러한 성능차이는 개념확장 대상 지식베이스를 선택함에 있어 최적의 지식베이스를 선정하지 못한 데서 나올 수 있다. 따라서 문헌 데이터베이스가 주제별로 분산 구축된 후 나올 수 있는 최적의 성능을 알아보기 위해 각 질문 별 최적의 성능이 나올 수 있는 주제범주를 수작업으로 추출하였고 (〈표 12〉 참조), 〈표 13〉은 1차 검색 결과의 분석을 통해 선택된 주제범주를 상위 10순위와 20순위로 나누어 보여준 것이다. 최적의 성능이 나올 수 있는 주제범주를 선택하는 작업은 초기에 질문을 작성하고 적합성 평가를 수행해 온 2명의 주제전문가에 의해 수행되었다.

표에서 보듯이 상위 10순위까지의 문헌을 분석하여 개념확장 대상 지식베이스를

〈표 13〉 1차 검색 결과의 분석을 통해 선택된 주제범주

평가문헌 수 \ 탐색문	1	2	3	4	5	6	7	8	9	10
10위	332	332	338	338	336	330	331	332	336	336
20위	332	332	332	338	336	330	332	331	336	336

〈표 14〉 최적 분산 DB 지식베이스 및 전체 DB기반 지식베이스간의 성능비교

검색성능 / 검색 문헌 수	재현율		정확률	
	KB분산_분산DB 검색(최적)	전체_DB	KB분산_분산DB 검색(최적)	전체_DB
10	0.1160	0.1333	0.5500	0.4700
20	0.1709	0.1947	0.4650	0.3900
30	0.2067	0.2437	0.3900	0.3433
50	0.2921	0.3350	0.3540	0.2780
평 균	0.1965	0.2267	0.4398	0.3703

선택할 경우 10개의 범주 중 최적의 범주와 일치하는 범주는 5개이고 20순위까지를 대상으로 할 경우 4개의 범주가 최적의 범주와 일치하는 것으로 나타났다.

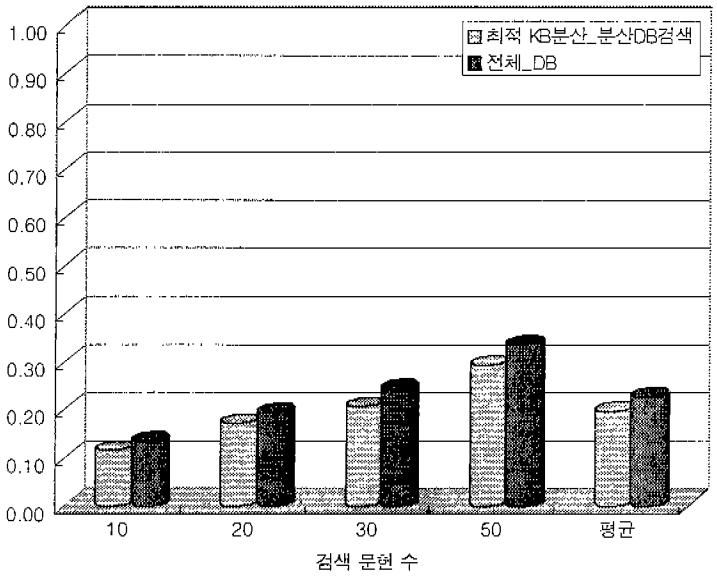
다음으로 1차 검색결과의 분석에 의해 개념확장 대상 지식베이스를 선택할 때 만약 최적의 범주만이 선택된다면 어느 정도의 성능이 나오는지 알아보고 전체 문헌 집단을 기반으로 지식베이스가 구축된 경우와 비교해 보았다. 〈표 14〉는 분산 구축된 지식베이스 중 최적의 분산 지식베이스를 선택했을 때와 전체 문헌을 기반으로 지식베이스를 구축했을 때의 재현율과 정확률을 보여 주고 있고 〈그림 6〉과 〈그림 7〉은 이를 그림으로 표현한 것이다.

〈표 14〉에서 보듯이 1차 검색결과에 의해 최적의 지식베이스를 선택했을 경우 전체 문헌 기반 지식베이스보다 재현율은 약 0.0302% 낮지만 정확률은 오히려 약 0.0695% 정도 높게 나타났다. 따라서 지식베이스를 분산 구축할 경우 최적의 지식베이스가 선택된다면 전체 문헌 데이터베이스를 기반으로 지식베이스를 구축하였을 때보다 더 높은 성능을 얻을 수 있음을 알 수 있다.

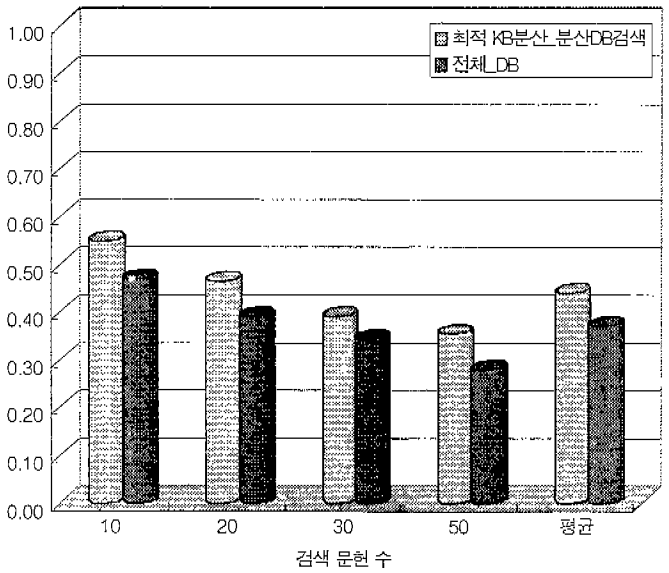
5 결 론

개념기반 정보검색기법이 불리언 검색기법의 단점을 어느 정도 해결한 것으로 평가되고 있는 단순 매칭함수 기법이나 P-norm 검색기법보다 높은 성능을 보여주고 있는 것으로 나타나고 있지만 개념확장을 위해 필수적인 의미망 지식베이스를 구축하는 시간이 너무 오래 걸리는 단점이 지적되어 왔다. 본 연구에서는 이러한 단점을 해결하기 위해 주제범주별로 지식베이스를 분산 구축함으로써 지식베이스 구축 시간을 줄이면서도 검색성능이 떨어지지 않도록 하는 방법을 모색하고자 하였다. 그 실험결과는 아래와 같다.

첫째, 주제범주별로 문헌을 분류한 후 각 범주별로 지식베이스를 분산 구축한 경우가 전체 문헌 데이터베이스를 대상으로 하나의 지식베이스를 구축한 경우보다 평균적으로 보았을 때, 약 8.6배 정도 지식베이스 구축 시간이 단축되는 것으로 나타났다.



〈그림 6〉 최적 분산 DB 지식베이스 및 전체 DB기반 지식베이스간의 재현율 비교



〈그림 7〉 최적 분산 DB 지식베이스 및 전체 DB기반 지식베이스간의 정확률 비교

둘째, 분산 구축된 지식베이스 중 적합한 개념확장 대상 지식베이스를 선정하는 방법 중 초기 탐색문을 가지고 전체 문헌 집단을 대상으로 1차 문헌 검색을 한 후 상위 순위에 있는 문헌들이 가장 많이 출현한 지식베이스를 선정하는 방법과 초기 탐색문으로 전체 문헌 집단을 대상으로 1차 검색을 한 후 상위 순위에 속한 문헌들의 유사도 가중치의 합이 가장 높은 지식베이스를 선정하는 방법을 비교한 결과 검색된 문헌들의 범주 별 유사도 가중치의 합으로 할 경우가 동순위 문제가 발생하지 않아 보다 적합한 것으로 판단되었다.

셋째, 분산 구축된 지식베이스 중 개념확장 대상 지식베이스를 선정하는 방법들 중 이용자가 제공한 초기 탐색어의 데이터베이스 내 문헌빈도에 확장용어 가중치를 반영한 경우가 가장 높은 성능을 보여주었다.

넷째, 전체 문헌 데이터베이스를 대상으로 지식베이스를 구축하고 이를 대상으로 검색한 경우가 가장 높은 성능을 보여주었고 지식베이스를 분산 구축한 경우의 성능은 개념확장을 하지 않고 단순 문헌 검색을 한 경우보다 약간 더 높게 나타났다.

마지막으로, 1차 검색결과의 분석에 의해 개념확장 대상 지식베이스를 선택할 때 만약 최적의 범주만이 선택된다면 전체 문헌기반 지식베이스보다 정확률을 높일 수 있는 것으로 나타났다.

본 연구의 실험결과를 통해 개념기반 정보검색시스템은 지식베이스의 분산 구축을 통해 가장 큰 문제로 대두되고 있는 지식베이스 구축 시간을 획기적으로 단축시킬 수 있다는 사실을 발견할 수 있었다. 즉, 약 4,500건을 대상으로 실험한 이번 실험에서 지식베이스 구축 시간을 거의 8.6배 정도 단축시킨 것으로 보아 데이터베이스가 몇 만 건으로 늘어날 경우 이 비율 및 효과는 훨씬 더 커질 것으로 예상된다. 검색성능에 있어서는 여러 범주 중 최적의 주제범주를 선택하는데 50%가 실패하고 두 번째 또는 세 번째로 적합한 범주를 택함으로써 전체 문헌 데이터베이스 기반 지식베이스를 기반으로 했을 때보다 재현율과 정확률이 각각 0.06%, 0.02%로 낮아 졌지만 큰 차이는 발생하지 않았다.

본 실험연구는 소규모 데이터베이스를 기반으로 하고 있다. 향후 보다 대규모의 데이터베이스를 기반으로 한 실험이 수행됨으로써 지식베이스의 분산 구축에 의한 시간단축 뿐만이 아니라 개념기반 검색을 위한 보다 효과적인 의미망 지식베이스 구축 방안 모색에 대한 연구가 진행되어야 할 것으로 보인다.

참 고 문 헌

노영희. 1999. 『의미망 지식베이스를 이용한 개념기반 정보검색기법에 관한 실험적 연구』. 박사학위논문.

연세대학교 대학원, 문헌정보학과.

노영희, 정영미. 2000. 의미망 지식베이스를 이용한 개념기반 정보검색기법에 관한 연구. 『정보관리학회지』, 17(3): 45-70.

노영희. 2000. 개념기반 검색을 위한 시소러스 관계의 효과적 활용방안에 관한 연구. 『정보관리학회지』, 17(4): 47-66.

박병수. 2000. 『UML을 이용한 지식베이스 기반 정보검색 시스템의 설계 및 구현』. 석사학위논문, 수원대학교 대학원, 전자계산학과.

Bookstein, A., and D. R. Swanson. 1975. "Probabilistic models for automatic indexing." *Journal of the American Society for Information Science*. 26(1): 45-50.

Chen, H., and V. Dhar. 1991. "Cognitive process as a basis for intelligent retrieval systems design." *Information Processing & Management*. 27(5): 405-432.

Chen, H., K. J. Lynch, K. Basu, and T. D. Ng. 1993. "Generating, integrating, and activating thesauri for concept-based document retrieval." *IEEE EXPERT, Special Series on Artificial Intelligence in Text-based Information Systems*. 8(2): 25-34.

Chen, H., P. Hau, R. Orwig, L. Hoopes, and J. F. Nunamaker. 1994. "Automatic concept classification of text from electronic meetings." *Communications of the ACM*. 37(10): 56-73.

Cohen, P. R., and R. Kjeldsen. 1987. "Information retrieval by constrained spreading activation in semantic networks." *Information Processing & Management*. 23 (4): 255-268.

Cohen, William W. and Yoram Singer. 1996. "Context-sensitive learning methods for text categorization." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 307-315.

Maron, M. E., and J. L. Kuhns. 1960. "On relevance, probabilistic indexing and information retrieval." *Journal of the ACM*. 7(3): 216-243.

Lewis, David D., Robert E. Schapire, James P. Callan, and Ron Papka. 1996. "Training algorithms for linear text classifiers." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*

(SIGIR), 298-306.

Moulinier, I., G. Raskinis, and J. Ganascia. 1996. "Text categorization: a symbolic approach." *In Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval.*

Salton, G., Edward A. Fox, and Harry Wu. 1983. "Extended Boolean information retrieval." *Communications of ACM*, 26(12): 1022-1036.

Shoval, P. 1985. "Principles, procedures and rules in an expert system for information retrieval." *Information Processing & Management*, 21(6): 475-487.

van Rijsbergen, C. J. 1979. *Information Retrieval.* London : Butterworths.